**Resumo “Llama 2: Open Foundation and Fine-Tuned Chat Models”**

1. Abstract

* Desenvolvimento do Llama 2: uma coleção de LLMs pré-treinadas e fine-tuned usando entre 7 biliões e 70 biliões de parâmetros
* Llama 2 – Chat: orientado a use cases de diálogo
* Os modelos do Llama 2 são melhores que os open-source, tendo em conta os benchmarks da META e da sua avaliação humana
* Neste artigo, irá ser mostrado todo o processo de fine-tuning e da melhoria da segurança para que se possa usar responsávelmente o Llama 2 – Chat.

1. Introduction

* LLMs são assistentes AI que dão suporte a tarefas que envolvem raciocínio. É usado em coding, escrita creativa, … . Fomentam a interação com os humanos através de uma interface de chat
* As LLMs são treinadas segundo a filosofia de Reinforcement Learning with Human Feedback (RLHF)
* Os custos computacionais requeridos para o desenvolvimento destes modelos influenciaram na sua melhoria

|  |  |  |
| --- | --- | --- |
| **Open pre-trained LLMs** | **Closed pre-trained LLMs** | **Closed product LLMs** |
| Bloom  Llama-1  Falcon | GPT-3  Chinchilla | ChatGPT  BARD  Claude |

* Estes closed product LLMs são muito fine-tuned de acordo com as preferências do utilizador, de forma a realçar a sua usabilidade e segurança. Este treino não é fácil de reproduzir, o que limita os progressos na investigação de LLMs
* A META desenvolveu o Llama-2, uma família de LLMs pré-treinadas e fine-tuned
* De forma geral, os modelos do Llama-2-Chat têm melhor desempenho que os modelos open-source; estão a par de alguns modelos mais fechados (closed-source), de acordo com as avaliações humanas efetuadas (são subjetivas)
* Além disso, a segurança foi aumentada nestes modelos através da anotação e tuning de dados muito específicos *(acredito que seja algo como “se disseres X, vais incentivar a um crime”)*
* LLAMA-2 é um upgrade de LLAMA-1 pois usa mais dados e com o dobro do tamanho
* Existem várias variantes tanto do LLAMA-2 como do LLAMA-2-CHAT
* LLAMA-2 tem algumas lacunas já identificadas pela META e cabe aos developers fazer testes de segurança e adaptar o modelo à aplicação a desenvolver.

1. Pretraining

* De forma a treinar o LLAMA-2, usou uma versão com mais performance de um auto-regressive tranformer
* Os dados do pré-treino são de fontes públicas mas não incluem informação sobre os produtos ou serviços da Meta, uma vez que contém dados pessoais de muita gente
* Detalhes do treino: pegaram na configuração e na arquitetura do LLAMA-1; usaram pré-normalização com RMSNorm; usaram a função de ativação SwiGLU; usaram como hiperparametros o optimizer AdamW, …; usaram o mesmo tokenizer que o do LLAMA-1 – o tamanho do vocabulário é de 32 mil tokens
* O treino aconteceu no Meta’s Research Super Cluster (RSC) e em clusters internos de produção – RSC usa NVIDIA Quantum InfiniBand e os clusters internos de produção usam RoCE (RDMA over converged Ethernet); …
* A avaliação dos modelos teve por base os seguinte benchmarks: coding, pensamento do senso comum (Commonsense Reasoning), conhecimento do mundo (World Knowledge), compreensão da escrita e matemática. Abaixo temos a comparação do LLAMA1 e LLAMA2 com os modelos open-source MPT e FALCON

Tabela

Descrição gerada automaticamente

* Em baixo temos a comparação do LLAMA-2 com alguns modelos closed-source GPT3.5, GPT4, PALM e PALM-2-L

Interface gráfica do usuário, Tabela

Descrição gerada automaticamente

1. Fine-tuning

* Foi usado um supervised fine-tuning. Pedes X, ele dá-te um X mais fraquinho e tu dizes mais detalhadamente aquilo que tu queres. Cada amostra consiste num prompt de pergunta/pedido e numa resposta
* Foi usado também o reinforcement learning with human feedback /RLHF
* Recolha de dados de preferências do humano: ver se realmente dão as respostas às suas perguntas e se dão respostas seguras, ou seja, se não te dizem, por exemplo, os passados detalhados para fazeres uma bomba. Foram dadas labels às respostas e o modelo aprendeu o que era bom ou mau de acordo com essas labels.
* Rewarding model: modelos que servem para dizer se o modelo está a ir bem ou se está a responder mal através de scores. Existe um modelo para a utilidade das respostas e outro para a segurança
* Em baixo temos uma tabela que nos diz a accuracy obtida por cada modelo usando alguns datasets
* Fine-tuning iterativo: dá-se um prompt, recolhe-se X respostas e escolhe-se a melhor. O modelo aprende a partir daí. Tabela

  Descrição gerada automaticamente

1. Segurança (Safety)

* Há que tomar algumas medidas de segurança de modo a que os modelos não deem informação capaz de causar o mal
* Foram usados 3 benchmarks para a segurança: veracidade, toxicidade e bias.

Tela de celular com publicação numa rede social

Descrição gerada automaticamente com confiança média

* Em baixo temos as performance em termos de veracidade e de toxicidade para 4 LLMs open-source.

Tabela

Descrição gerada automaticamente

* Categorias de segurança: atividades ilícitas ou criminais; atividades de ódio; aconselhamento indevido (e.g. aconselhamento médico, financeiro, …)
* Utilidade do RLHF:

Texto

Descrição gerada automaticamente

* Em baixo temos as performances para modelos closed-source

Tabela

Descrição gerada automaticamente

1. Discussão

* LLMs contêm limitações como a perceção temporal – por exemplo, o chatgpt só tem conhecimento até 2021
* Secção 5.3: contém link para usar o LLAMA2: <https://ai.meta.com/resources/models-and-libraries/llama/>
* Amostras de código para se usar com o LLAMA2: <https://github.com/facebookresearch/llama>

1. Conclusão

* Resumo do artigo
* LLAMA-2 e LLAMA2-CHAT estã abertos ao público

1. Apendice

* Exemplos de respostas de várias LLMs a um dado prompt
* Explicações mais detalhadas a pormenores do fine-tuning, training, …